A central repository of publication results, implemented as a part of systems for revealing plagiarism.

Daniel Jakubík, Ľuboš Lunter, Michal Brandejs, Jitka Brandejsová

Faculty of Informatics, Masaryk University Botanická 68a, 602 00 Brno, Czech Republic
jakubik@fi.muni.cz; lunter@fi.muni.cz; brandejs@fi.muni.cz; brandejsova@fi.muni.cz

Abstract: Repozitar.cz is a new system for detection of plagiarism that will join the systems like Theses.cz and Odevzdej.cz. The system will provide the services of a scientific digital library. This project is an effort of 15 universities which with the technical solution form the necessary organizational, social and legal environment. This contribution introduces this system whose aim is, among others, the presentation of works according to Open Access idea, searching for similar documents or transfer of publication meta data to the RIV.

1. Introduction

As the first university in the Czech Republic, Masaryk University signed the Berlin Declaration in October 2010. This step commits the University to actively support and promote open access ideas (i.e. open access to scientific information). At the same time the University has incorporated support for Open Access to its long-term plan for the years 2011 - 2015. Based on these decisions, the development of the first version of the university repository commenced as a part of open services of the university Information System.

In 2011, the Faculty of Informatics MU started work on a project focusing on creating an inter-university network of technical and methodological tools and strategies for fighting plagiarism. As part of the project, Repozitar.cz was built to operate as the repository for storage and presentation of scientific works and technical papers.. The aforementioned repository in use at Masaryk University was chosen as the base for the system.

The following text describes both the steps already implemented and the remaining tasks outstanding.

2. Motivation

There were several incentives that lead to the development of the system Repozitar.cz.

- **Commitment to Open Access.**
  Access to previous results and their reuse in new research are at the very basis of scientific progress. These essential principles are the fundamentals that form the Open Access paradigm. This stands on the idea that the scientific works, especially peer-reviewed articles, should be provided on the internet for free.

Generally speaking, there are two different possibilities for Open Access publishing: the gold road and the green road. The gold road is essentially publishing in open access journals which are available to the reader online and free of charge. The green road stands on self-archiving, which is to say that the authors deposit their works in institutional or subject-bound repositories. Development of institutional repositories is the main way that the Open Access ideal is progressed by the universities.[1]
As a signatory to the Berlin Declaration, the Masaryk University has assumed responsibility for the active dissemination of knowledge; not only through the traditional form but also increasingly through the Open Access paradigm. For this reason, the Repozitar.cz system will provide open access repository services under the Creative Commons licence.

- **Evidence and long-term archiving of scientific publications.**
  Scientific data in digital formats have become an integral part of university production. There is a slight paradox in that digital media and data is not always accessible, while for the university, the informational value lies in its dissemination to interested parties. Preservation of data for future generations is a major priority, and so one of the challenges for the universities will be to guarantee the long-term archiving of important scientific data and research-related information.

- **Plagiarism detection in scientific outlets.**
  The problems with plagiarism detection have been dealt with by Masaryk University since 2006. At that time users of the Information System of Masaryk University (IS MU) could start using the functions of tracing similarities (potential plagiarisms) - not only in theses and e-learning materials but also in all documents stored in the system.

  In 2008, a joint initiative of most public universities in the Czech Republic resulted in the development of a plagiarism detection system - Theses.cz - which is currently used by 35 universities (including 2 from Slovakia). This was followed by the implementation of a system detecting plagiarisms in seminar papers called Odevzdej.cz in 2009. Odevzdej.cz has been used recently by 27 universities, some secondary schools and the Ministry of Labour, Social Affairs and Family of the Slovak Republic.

  The project Odevzdej.cz came was developed as a reaction to the teachers' demand for the detection of plagiarism in theses. The rationale was that at humanities oriented schools seminar papers and other works represent an indispensable part of student's studies. A large number of Czech universities are also interested in continuing with an interuniversity control of plagiarisms in the field of scientific and professional publications. Apart from striving to prevent plagiarism, they also aim to eliminate the negative effect of servers collecting academic works and enabling students to copy them.

  Last but not least, all three systems of plagiarism detection set out to educate academic society in how to exploit information sources they use when writing papers.

- **Submitting to RIV.**

  Registry of R&D results (RIV) is collecting and containing information about the results of research projects financed from public budgets. Correct transfer of data to its central database is one of the basic conditions for the future granting of institutional resources University research and development.

Users of the Information System of Masaryk University have at their disposal a publication agenda enabling continuous data collection. Every year, multiple incremental outputs are automatically generated from the data collected during the previous year. A significant number of bugs and incorrect records were eliminated through an extensive system of controls. Therefore, in recent years, the only major errors left are conflicts with other institutions, for example when an author with multiple workloads being marked as local by two or more institutions.

These types of errors cannot be controlled within a single institution. To solve this problem it is necessary to aggregate data from other institutions as well.

3. Building the repository

As mentioned above, the system Repozitar.cz rises from the platform of the Masaryk University repository. Therefore it is necessary to describe that system first:

Based on extensive discussions about the concept of the institutional repository of Masaryk University the following conclusions emerged:[2]

1. The repository has to be designed with a specific goal of minimising the burden on the authors. Their administrative load is already too high, so its further increase raises the risk of refusals to submit their works to the repository.

2. The authors should benefit from depositing of their full-texts in the Repository so that they are further motivated to use it.

3. The Repository should be conceived from the start to interconnect with other University systems and activities to achieve synergy effects and to minimize its costs.

Two options for the implementation of the repository were considered: (1) developing a new system and (2) integration into the Information System of the Masaryk University. Taking the three criteria into account, the latter option was chosen.

Several parts of the system were developed as mutually independent but with later possible integration being a key concern.

- **Module for inserting of the new publication records.**
  This module provides functionality for adding new entries to the repository. The registry of publications and results of scientific activities in the IS MU has become the basis for the module. It consists of a set of web forms that were derived from the RIV rules definition. Each form serves for the registration of publication records of various types, which differ by set of collected data. The module has been extended with an option to store any file e.g. pre-prints, post-prints or scientific data. So the primary publication records represent the meta-data which are closely bound to any number of attached files.

- **Data deposit.**
  All the attached files are stored in the Deposit which uses the same technology as Theses.cz or Odevzdej.cz. It gives access to the advanced functions as automatic scanning of documents or the plagiarism detection system. The Deposit possesses a wide range of access rights which can be combined. It allows to the authors to create different levels of permissions for different users for each file, with respect to their agreement with publisher. Moreover, the information about the type and chosen open access licence is registered for every attached file.

- **Access interface.**
  The main access interface is based on a full-text search system developed for searching within large collections of documents. The system creates indexes and consequently can search for meta-data as well as the full text of the articles and other publication records. Because there is functionality to add any additional information in the form of “virtual tokens” to the index, the search system can be used to generate automatically various lists, for example a list of publications published in the same journal.

  The advanced search options extend the application by adding the possibility of progressive refinements to the query. It allows search not only by publication meta-data but also by departments, R&D projects or other data required for transmission to the RIV.

  A final list of matching records can be entered into a user box which enables the transfer of a selection into other applications, work with it and subsequently process them en bloc.

  The system also allows the transfer of its content to other systems via web-based protocol OAI-PMH Version 2.0. The scope of the data transferred and their recipient depend only on the agreement among the participating universities and the individual institutes of the Academy of Sciences who serve as the data providers.

- **Layer of related applications**
  A set of supporting applications were developed around the core of the repository. A typical example is a utility for the presentation and printing of publication records via user templates. The combination of dedicated applications and a universal mechanism for transferring search results permits the expansion of the application layer easily by adding new services.

  The main goal of the system Repozitar.cz is to provide a solution for centralized gathering, controlled presentation and checking against plagiarism in the scientific papers, publications and other works produced by research workers and doctoral students. The system is conceived as a central system with decentralized services for each school or institution. The central part of the system provides functionality for accounts administration, harvesting data, controlled presentation and plagiarism detection of full-text data. For the individual institutions the most important aspect is to provide the services for transferring the data from their local systems to the central repository.

4. Import of publication records to the repository

The high number of participating schools and institutes puts complex demands on openness of the system for different types of requirements. The publication records can be inserted into the system by individual authors as well as by different automated methods. A simple import of the works is realized via a web form into which the meta-data can be filled and consequently saved. Another option is to transmit the data using the program Curl or through the OAI-PMH protocol.

The repository supports several bibliography formats as BibTeX, RIS, MARC or Dublin Core which can be used to import meta-data into the system. In addition, the repository defines its own format based on XML which reflects the internal database structure. The format consists of meta-data collected by RIV and additional information for repository purposes. The file-path information is also attached as a part of the meta-data format. These files are automatically downloaded and the embedded OCR system recognizes this text information.

All files entered into the system are scanned by an anti-virus and periodically backed up.

5. Access to the content

From the author's perspective the Repozitar.cz system appears to be a web-based bookcase into which the authors can submit any type of multimedia data, bind them with their meta-data and consequently work with them (e.g. to register the records which cite their work). However, not all submitted records must be publicly accessible. It depends on the data providers' selections as to which outputs they want to present for themselves and their scientific research activity.

At Masaryk University, the dissemination of academic results is based on the assumption that the author knows best about the rights of the third parties involved and how their work can be dealt with. The first step is to determine the nature of the work – if it is job-related or not. Work funded by grants and transmitted to the RIV, i.e. the works bound to some project, are usually job-related and so the employer maintains their proprietary rights. The large volume of articles at MU complicated the settlement of the accessibility for each article in proceedings or periodicals by MU itself, therefore, by a special directive, the university abnegated the performing of proprietary rights to these types of works. Yet because these works were created in a relation to a MU occupation[3] other internal directive imposes a duty of the authors to submit the record into the repository and to set up its access rights. When saving other types of job-related publications than those from proceedings or periodicals, an employee is entitled to suggest the extent of its disclosure which has to be approved by a dean or by his delegate.

Because the repository can hold different versions of the same work with different access rights, it is possible to comply with requirements of publishers who often only allow the making accessible of earlier (e.g. submitted or accepted) versions of the publication. The repository allows setting up deferred releasing of works for public too. In the case that the full text of the work is not available at all or only a part is available, it will be possible to search the records via its meta-data and to request the full text directly from the author.

6. Conclusion

his paper provides insights into the implementation of a new system - Repozitar.cz - which arises through the mutual cooperation of 15 universities. The system provides services for long-term storage and presentation of academic publications. According to the Open Access idea there is an effort to make the most of records open to the public, but for various reasons it is not possible to fully achieve this goal. Therefore authors can use a wide range of access rights to restrict the access to their publication outputs. Since a lot of repositories struggled with little interest from the authors submitting data into them, Repozitar.cz tries to attract them by facilitating their duty to submit publication records into the RIV. For this reason the set of data collected to RIV is a subset of data collected by the repository and the system enables automatic submission of the records into this RIV register. In addition to these services, Repozitar.cz includes the same detection for plagiarism as systems Theses.cz and Odevzdej.cz and its content enlarges common databases of all three systems. For the future, many tasks remain; nevertheless we can already see some great results that we have achieved so far. Together our team expects that Repozitar.cz will play an important role in the future of digital libraries in the Czech Republic, and we look forward to the productive collaboration with other institutions in this area.

References

[1] Peter Suber. Research on institutional repositories: Articles and presentations. *Open Access Overview*, paper 45, 2010. URL `http://digitalcommons.bepress.com/repository-research/45`.

[2] Miroslav Bartošek and Michal Brandejs and Ivana Černá. Otevřený přístup k vědeckým informacím na masarykově univerzitě. *Zpravodaj ÚVT MU*, XXI(5):1, 2011. ISSN 1212-0901. URL `http://www.ics.muni.cz/zpravodaj/articles/673.html`.

[3] Úplné znění zákona č. 121/2000 sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), jak vyplývá z pozdějších změn, 2000. URL `http://www.sbirkazakonu.info/autorsky-zakon/`. §58, odstavec 1.